THE BEHRENS-FISHER PROBLEM; A COMPARISON OF TEST PROPERTIES WHEN OBSERVATIONS ARE NOT NECESSARILY NORMAL

Grete U. Fenstad Eva Skovlund

Short title:

The Behrens-Fisher problem; a comparison of test properties

Corresponding address:

Department of Mathematics
University of Oslo
P.O.Box 1053 Blindern
0316 Oslo, Norway

THE BEHRENS-FISHER PROBLEM A COMPARISON OF TEST PROPERTIES WHEN OBSERVATIONS ARE NOT NECESSARILY NORMAL

Grete U. Fenstad Eva Skovlund

November 1992

ABSTRACT

The Behrens-Fisher problem is the problem of testing the hypothesis of equality of the means of two normal distributions with possibly unequal variances. Three tests to take care of this problem have been developed by Welch and Scheffé. The properties of these tests are compared with the properties of the ordinary T test and the Wilcoxon test both asymptotically and by stochastic simulation. In addition to considering the normal situation, which has been studied previously, simulations are performed with observations generated from a number of distributions — both symmetric and with heavy right tails. We find that Welch' tests have best overall performance.

KEY WORDS

Behrens-Fisher problem, hypothesis test, unequal variances, significance level, asymptotic properties, stochastic simulation.

1 INTRODUCTION

Testing the hypothesis of equality of the means of two normal distributions with unknown unequal variances is called the Behrens-Fisher problem. More precisely consider the two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from $N(\xi, \sigma^2)$ and $N(\eta, \tau^2)$ respectively. The problem is that of testing the hypothesis H: $\xi = \eta$ against the alternative A: $\xi \neq \eta$.

We introduce the statistics $\bar{X} = \sum_{i=1}^{m} X_i/m$, $\bar{Y} = \sum_{j=1}^{n} Y_j/n$, $Z_1 = \sum_{i=1}^{m}(X_i - \bar{X})^2$, $Z_2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2$, and the parameter $\theta = \sigma^2/\tau^2$.

When $\theta = 1$, it is well known that the test based on

$$T = (\bar{X} - \bar{Y})/\sqrt{(\frac{1}{m} + \frac{1}{n})(Z_1 + Z_2)/(m + n - 2)}$$

is uniformly most powerful among unbiased size ϵ tests. If the same test is applied when $\theta \neq 1$, the significance level may be far from the nominal level ϵ. In this case other tests should be applied.

Fenstad (1983) has compared tests based on

$$U = (\bar{X} - \bar{Y})/\sqrt{\frac{Z_1}{m(m-1)} + \frac{Z_2}{n(n-1)}}$$

and

$$V = (\bar{X} - \bar{Y})/\sqrt{\frac{Z_1}{m(m-3)} + \frac{Z_2}{n(n-3)}}$$

for varying θ. The test statistics U and V have been suggested by Welch (1937). His reasoning behind the V test is to find a test statistic with a variance which varies as little as possible with θ. For fixed critical values in particular, a test based on V has a more stable significance level than a test based on U, see Fenstad (1983).

Another test recommended in many textbooks is the test introduced by Scheffé (1943). Assuming $m \leq n$, let

$$S_i = X_i - \sqrt{\frac{m}{n}} Y_i + \sqrt{\frac{m}{n}} \frac{1}{m} \sum_{j=1}^{m} Y_j - \bar{Y}, \qquad i = 1, \ldots, m.$$

$S_1, \ldots, S_m$ are i. i. d. $N(\xi - \eta, \sigma^2 + \frac{m}{n}\tau^2)$ and the statistic

$$S = \frac{\bar{S}\sqrt{m}}{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(S_i - \bar{S})^2}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(S_i - \bar{S})^2}}\sqrt{m}$$

has a t-distribution with $m - 1$ degrees of freedom under the hypothesis. A test based on S will thus have a significance level which is independent of θ.

We have in this paper removed the normal assumptions and have considered the situation where F and G are continuous distribution functions with means ξ and η and finite fourth order moments. We wanted to test the hypothesis H: $\xi = \eta$ against the alternative A: $\xi \neq \eta$ on basis of the two samples $X_1, \ldots, X_m$ from F and $Y_1, \ldots, Y_n$ from G. It was

then natural also to include the Wilcoxon test based on the Mann-Whitney statistic W = number of pairs (X_i, Y_j) with $X_i < Y_j$.

If *(i)* F and G are normal distributions with common variance, the T statistic is used; if *(ii)* G is a pure translation of F, the W statistic is the relevant statistic; and if *(iii)* F and G are normal distributions with possibly different variances, the statistics S, U, and V come into consideration. We have compared tests based on T, U, V, S, and W for some specific choices of F and G covering these situations and also the situation when *(iv)* F and G are members of the same location-scale family with finite fourth order moment, not necessarily normal.

A comparison of tests based on the T, U, and V statistics in the situations *(i)* and *(iii)* has been performed by Fenstad (1983) who recommends the V test. Best and Rayner (1987) disagree with this conclusion; however, their simulations show that the U and V tests alternate in closeness to the the nominal level. We therefore have repeated the simulations for the normal case (situations *(i)* and *(iii)*) and have added simulations of other distributions (situations *(ii)* and *(iv)*).

As we also wanted to consider the situations *(ii)* and *(iv)* we brought in the statistic W to take care of *(ii)*. The significance level and power function of a test based on S is well known in the situations *(i)* and *(iii)*. Even if this test has a significance level independent of θ there seems to be a reluctance to use it in practice, since the test result might depend on the permutation of the Y's (see for example Scheffé (1970)). However, we wanted to see how it compared with the other tests in situations *(ii)* and *(iv)* with respect to significance level as well as power function.

2 DEFINITION OF THE TESTS

To decide for which values of the test statistic to reject the hypothesis, we considered the distributions of the test statistics under the hypothesis. The T statistic is designed to take care of situation *(i)* and then has a t-distribution with $m + n - 2$ degrees of freedom. A test based on T has significance level ϵ if we reject when

$$|T| > t_{\epsilon/2;m+n-2}$$

where $t_{\alpha;f}$ is the upper α fractile of the t-distribution with f degrees of freedom.

The S statistic has a t-distribution with $m - 1$ degrees of freedom in situations *(i)* and *(iii)*. The test which rejects when

$$|S| > t_{\epsilon/2;m-1}$$

will have significance level ϵ.

Since the distributions of U and V cannot be given a closed form, one uses the fact that they both are approximately distributed as $t_f/\sqrt{c}$, where c is a constant and t_f is t-distributed with f degrees of freedom. For U $c_U = 1$ and

$$f_U = f_U(\theta) = \frac{(\theta/m + 1/n)^2}{\theta^2/m^2(m-1) + 1/n^2(n-1)}$$

and for V

$$c_V = c_V(\theta) = \frac{(m-1)\theta/m(m-3) + (n-1)/n(n-3)}{\theta/m + 1/n}$$

and

$$f_V = f_V(\theta) = \frac{[(m-1)\theta/m(m-3) + (n-1)/n(n-3)]^2}{(m-1)\theta^2/m^2(m-3)^2 + (n-1)/n^2(n-3)^2}$$

Finally the critical values are obtained by plugging in an estimate for θ,

$$\hat{\theta} = \frac{Z_1/(m-1)}{Z_2/(n-1)}.$$

Thus the test based on U rejects H if

$$|U| > t_{\epsilon/2;f_U(\hat{\theta})}$$

and the test based on V rejects H if

$$|V| > t_{\epsilon/2;f_V(\hat{\theta})}(c_{V(\hat{\theta})})^{-1/2}.$$

These tests have both an approximate significance level ϵ in situations *(i)* and *(iii)*.

Finally, the distribution of the Mann-Whitney statistic W is known in situations *(i)* and *(ii)*, and its critical values are tabulated for moderately small values of m and n. For larger m and n we use that W is approximately $N(mn/2, mn(m+n+1)/12)$ in these situations and reject H if

$$|W - mn/2| > z_{\epsilon/2}\sqrt{mn(m+n+1)/12}$$

where z_α is the upper α fractile in the standard normal distribution. To obtain exact significance level ϵ we have randomized when necessary.

3 ASYMPTOTIC PROPERTIES

The five tests in consideration have either exact or approximate significance level ϵ in situation *(i)*. In this section we will examine the asymptotic significance levels as $m, n \to \infty$, $m/n \to \nu$ for the five tests in the other situations. The parameter $\theta = \sigma^2/\tau^2$ where σ^2 and τ^2 are the variances of F and G, respectively.

Since $Z_1/m \xrightarrow{P} \sigma^2$ and $Z_2/n \xrightarrow{P} \tau^2$, it easily follows from the central limit theorem that the asymptotic distribution of T is $N(0, \frac{\theta+\nu}{\theta\nu+1})$. Since $t_{\epsilon/2;m+n-2} \to z_{\epsilon/2}$, the asymptotic significance level of the T test is

$$\epsilon_T(\theta) = 2[1 - \Phi(z_{\epsilon/2}R(\nu,\theta))]$$

where Φ is the standard normal distribution function and $R^2(\nu,\theta) = (\theta\nu+1)/(\theta+\nu)$.

The S test has exactly significance level ϵ in situations *(i)* and *(iii)* when the observations are normal. In general $(\bar{X}-\bar{Y})/\sqrt{\sigma^2/m + \tau^2/n} \xrightarrow{d} N(0,1)$, and it remains to consider

the denominator $D = \sum_{i=1}^{m}(S_i - \bar{S})^2/(m-1)$. We write $\sum_{i=1}^{m}(S_i - \bar{S})^2 = \sum_{i=1}^{m}(R_i - \bar{R})^2$ where $R_i = (X_i - \xi) - \sqrt{(m/n)}(Y_i - \eta), i = 1, 2, \ldots, m$ are i.i.d. After some calculation we find

$$\mathrm{Var}(D) = \frac{1}{m}[\lambda_{4,X} + 6\frac{m}{n}\sigma^2\tau^2 + (\frac{m}{n})^2\lambda_{4,Y}] - \frac{1}{m}\frac{m-3}{m-1}(\sigma^2 + \frac{m}{n}\tau^2)^2 \to 0,$$

where $\lambda_{4,X} = \mathrm{E}(X_i - \xi)^4$ and $\lambda_{4,Y} = \mathrm{E}(Y_i - \eta)^4$. Since $\mathrm{E}(D) = \sigma^2 + (m/n)\tau^2$ a generalization of Chebyshev's inequality implies

$$D = \frac{1}{m-1}\sum_{i=1}^{m}(S_i - \bar{S})^2 \xrightarrow{P} \sigma^2 + \nu\tau^2.$$

Hence the test statistic S has asymptotic standard normal distribution and asymptotic significance level ϵ also in cases *(ii)* and *(iv)*.

The statistics U and V will both have asymptotic standard normal distributions. Since

$$t_{\epsilon/2;f_U(\hat{\theta})} \xrightarrow{P} z_{\epsilon/2} \quad \text{and} \quad t_{\epsilon/2;f_V(\hat{\theta})}(c_V(\hat{\theta}))^{-1/2} \xrightarrow{P} z_{\epsilon/2}$$

as $m, n \to \infty, m/n \to \nu$, the two tests based on U and V respectively will have asymptotic significance levels ϵ.

Finally we consider the asymptotic properties of W. Let

$$\begin{aligned} p_1(\theta) &= \mathrm{P}(X < Y) \\ p_2(\theta) &= \mathrm{P}(X < Y, X < Y') \\ p_3(\theta) &= \mathrm{P}(X < Y, X' < Y) \end{aligned}$$

where X, X', Y, Y' are independent random variables, X, X' with distribution F and Y, Y' with distribution G. It can be shown (Lehmann (1975), Example 20 in Appendix) that $\mathrm{E}(W) = mnp_1(\theta)$ and

$$\mathrm{Var}(W) = mnp_1(\theta)(1 - p_1(\theta)) + mn(n-1)(p_2(\theta) - p_1(\theta)^2) + mn(m-1)(p_3(\theta) - p_1(\theta)^2)$$

and that $(W - \mathrm{E}(W))/\sqrt{\mathrm{Var}(W)}$ has an asymptotic standard normal distribution. In cases *(i)* and *(ii)* $\mathrm{E}_H(W) = mn/2$ and the variance expression reduces to $\mathrm{Var}_H(W) = mn(m+n+1)/12$, hence the asymptotic significance levels are ϵ.

In general, if $p_1(\theta) = 1/2$, as is the case if the distribution functions are symmetric, the asymptotic significance level is

$$\epsilon_W(\theta) = 2[1 - \Phi(z_{\epsilon/2}R(\nu, \theta))]$$

where

$$R^2(\nu, \theta) = \frac{1}{12}\frac{\nu + 1}{p_2(\theta) - 1/4 + \nu(p_3(\theta) - 1/4)}$$

If $p_1(\theta) \neq \frac{1}{2}$, we obtain an asymptotic significance level equal to 1, and the Wilcoxon test should therefore not be used in this case.

To summarize, all five tests considered here have exact or asymptotic significance levels ϵ in cases *(i)* and *(ii)*. In cases *(iii)* and *(iv)* the U, V, and S tests also have asymptotic significance level ϵ, while the T and W tests have significance levels which vary with θ as well as ν. The asymptotic significance level of the W test varies even with F and G.

We give explicit expressions for $p_2(\theta)$ and $p_3(\theta)$ in the situations we simulate in the Appendix.

4 SIMULATION

The small sample properties of the four tests were explored by stochastic simulation. The simulation programs were written in SIMULA (Birtwistle et al, 1983). The pseudo-random number generator in SIMULA is a multiplicative congruential generator (Bratley et al, 1983). The simulations were executed on a SUN work station at the University of Oslo.

In the simulation program independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ were drawn from distributions with the same shape and mean, but possibly different variance. The ratio between the variances in the two groups, θ, was varied between 1/16 and 16.

The estimated number of degrees of freedom for the U and V tests is usually not an integer. The fractiles $t_{\epsilon/2,f}$ were determined by (Wang, 1971)

$$t^2_{\epsilon/2,f} = f \cdot (exp(z^2_{\epsilon/2}/g(f)) - 1)$$

where $g(f) = 0.9975 \cdot f - 0.445$ and $z_{\epsilon/2}$ is the upper $\epsilon/2$ fractile of the standard normal distribution.

The simulations were performed over a range of different distributions — symmetric (normal, uniform and double exponential) and with a heavy right tail (gamma and exponential). The simulations were mostly executed with equal sample sizes ($m = n = 10$) and with unequal sample sizes ($m = 5$, $n = 15$). The parameters of the distributions we simulate will be defined in the Appendix.

To estimate the actual significance level of each test, and in a few cases the power function, $N = 10000$ simulations were performed in order to obtain acceptable confidence interval lengths. (Normal approximation to the binomial distribution then gives an interval length less than 0.0196 when the coefficient of confidence is 0.95.) All four tests were compared in the same simulation run (i.e. with identical samples) in order to remove stochastic variation between the test results.

5 RESULTS

5.1 Comparison of the U and V tests

To compare the significance levels of the U and V tests when $m \neq n$ in the situations we consider, we estimated the significance levels based on 10.000 simulations. The standard deviation of the estimates is less than .005. The results are shown in Table 1.

	Normal		Uniform		Double exp.		Exponential		Gamma ($r = 2$)	
θ	U	V	U	V	U	V	U	V	U	V
1/16	.0511	.0504	.0588	.0587	.0449	.0438	.0623	.0594	.0541	.0527
1/9	.0497	.0481	.0557	.0547	.0445	.0419	.0494	.0450	.0452	.0426
1/4	.0512	.0460	.0525	.0484	.0416	.0377	.0374	.0352	.0400	.0375
1	.0565	.0480	.0643	.0567	.0390	.0330	.0710	.0678	.0644	.0583
4	.0570	.0513	.0702	.0649	.0393	.0343	.1166	.1081	.0862	.0783
9	.0524	.0490	.0695	.0662	.0372	.0326	.1220	.1147	.0877	.0819
16	.0504	.0477	.0685	.0663	.0354	.0324	.1224	.1171	.0866	.0824

Table 1 Estimated values of ϵ for the U and V tests when $m = 5$ and $n = 15$, based on $N = 10000$ simulations. When $m = n$ the tests are identical. The nominal level is $\epsilon = 0.05$.

We conclude that the V test is closer to the nominal level than the U test for most of the points in the cases of uniform and gamma ($r = 2$) distributions, while in the normal case the result is more uncertain. In the exponential and the double exponential cases both tests are too far from the nominal level. For most of the chosen θ values the V test is more conservative than the U test and therefore performs better than the U test when the significance level is too high and worse for too low significance levels. This result is specific for the chosen values of $m = 5$ and $n = 15$, for instance one obtains the opposite result if $m = 15$ and $n = 5$.

5.2 Sample sizes equal ($m = n$)

Figure 1 shows the simulated significance levels for the T test and the W test for different sample sizes and different values of θ compared with the asymptotic significance levels for $m = n$ when the observations are normal. Unlike the T test the W test does not have asymptotic significance level ϵ. It is also seen that the significance level of the T test improves as $m, n \to \infty$, whereas the significance level of the W test gets worse as $m, n \to \infty$.

The S test has significance level ϵ independent of θ in the normal case, while the U and V tests have asymptotic significance level ϵ.

When sample sizes are equal, the U and V tests are identical, we therefore only refer to the V test. Simulations show that the significance level of the V test is close to the nominal level in the normal case. When observations are uniform, the simulated significance levels for the T, W, S, and V tests are close to the results for the normal case. When observations are drawn from double exponential distributions, the simulation results are also close to the normal case, but with slightly lower levels for the S, T, and V tests.

To represent simulation results when distributions are skew, the gamma distribution

with parameter $r = 2$ was chosen. Figure 2 shows simulated significance levels of the S, T, V, and W tests for varying θ when $m = n = 10$ and the observations come from a gamma distribution. All tests obtain significance levels which are too high when $\theta \neq 1$. The simulated level of the W test is considerably higher than the levels of the other three, which is in agreement with the asymptotic results. Generating observations from gamma distributions with $r = 3$ (less skew), gives results which are closer to the nominal level.

When observations are generated from the exponential distribution, the simulated levels for the S, T, and V tests increase only slightly, whereas the level for the W test increases substantially when distributions are more skew and $\theta \neq 1$. For instance, when the observations are exponential and $\theta = 1/9$, the simulated level is 0.159.

5.3 Sample sizes unequal ($m = 5$, $n = 15$)

When the sample sizes m and n are unequal, the significance levels are no longer symmetrical around $\theta = 1$. Figure 3 shows simulated significance levels when $m = 5$ and $n = 15$ and the observations are normal. The S test has significance level ϵ independent of θ, and the V test has approximately so. The simulations show, however, that the T and W tests have significance levels which deviate substantially from the nominal level. When the smaller sample is drawn from the distribution with the smaller variance, the actual level is too low, whereas it is too high when the larger sample comes from this distribution.

When the observations are uniform or double exponential, the results for the T and W test are largely similar to the results for the normal case. The S and V tests have almost identical levels when the observations are uniform; increasing from about 0.045 when θ is small to slightly below 0.07 when θ is large. When the observations are double exponentially distributed, the significance level of the S test changes from approximately 0.06 to 0.04 with increasing θ, whereas the level of the V test changes gradually from 0.045 to 0.035. Hence, the results for these two tests deviate only slightly from the nominal level compared to the T and W tests when the observations come from symmetric but non-normal distributions.

The gamma distribution with parameter $r = 2$ was again chosen to represent the simulation results when the observations are generated from skew distributions. Figure 4 shows that the simulated significance levels of all four tests deviated from the nominal level, but as for the symmetric distributions, the results for the S and V tests are considerably closer to the nominal level than the other two. In contrast to the situation when $m = n = 10$, the level of the W test is closer to the nominal level than the level of the T test.

Generating observations from the less skew gamma distributions gives simulated levels closer to the nominal, whereas generating observations from the exponential distribution gives larger deviations.

5.4 Power functions

When $\theta = 1$ the difference between the powers of the T and V tests is at most about 10% when $m = 5, n = 15$. The significance level of the T test changes rapidly with θ and power functions are therefore not of interest when $\theta \neq 1$.

Since the S test and the V test obtain significance levels which are mostly quite similar and close to the nominal level, it was of interest also to compare their power functions. Simulations showed that the V test has higher power than the S test for all cases examined. Figure 5 shows power functions for the two tests when the observations are normal, $\theta = 1/9$, and the sample sizes are $m = 5$ and $n = 15$. The V test has considerably higher power than the S test. When $\theta = 1$ the power of the V test is a little lower than in the previous case, whereas the power function for the S test is unchanged. When the sample sizes are equal ($m = n = 10$), the power functions for both tests are close to the power function of the V test shown in the figure.

6 DISCUSSION

The properties of five tests have been compared. The T and W tests were chosen because they are used extensively. The U test is also much in use, while the S and V tests are less often used for practical purposes, although they have been developed especially for the Behrens-Fisher situation.

The observations have been generated from a set of distributions with different properties. The normal distribution was included as a "standard" although some of the test properties are known in the normal case. Two other symmetric distributions have also been included for generation of observations, the uniform distribution which has smaller kurtosis than the normal distribution, and the double exponential distribution which has larger kurtosis than the normal distribution.

Skew distributions have also been considered; the gamma distribution with three different choices of parameters, the exponential distribution being one of them.

Many other distributions could have been included in this simulation study, but in our opinion this choice of different symmetric and skew distributions provides results which are representative for a range of distributions with different skewness and kurtosis.

We believe gamma distributions often fit actual empirical distributions well, for instance data from biology or medicine. Skovlund and Walløe (1991) have fitted gamma distributions to patient data from a clinical trial on myocardial infarction. This example also illustrates a case where θ may be different from 1; the group with the larger mean also tends to have a larger variance. Therefore it seemed natural to examine test properties when observations are generated from gamma distributions. Observations have been generated from gamma distributions with $r = 2$ and $r = 3$. The gamma distribution with $r = 2$ has been chosen to draw the figures since this distribution is more skew than the distribution with $r = 3$.

The gamma distribution has two parameters r and λ. Since skewness and kurtosis in this distribution are determined only by r, we have chosen to keep r fixed and vary λ in order to change the variance.

We have shown that the properties of the S, U, and V tests are considerably better than the properties of the T and W tests in the Behrens-Fisher situation, and that the T and W tests should not be used if it is not known that $\theta = 1$. Both the S test and the U and V tests obtain significance levels which are mostly very close to the nominal level

also for a range of non-normal distributions. Since the performance of the U and V tests are not very different, the U test has been deleted in the figures. The U and V tests have higher power than the S test, and do not have any "permutation problems". Of the U and V tests we recommend the V test as the better choice when testing the equality of the expectations of two distributions with possibly unequal variances, since it performs better in the normal situations, *(i)* and *(iii)*.

APPENDIX

Observations were generated from the following distributions, where we give the densities explicitly to identify the parameters:

Normal: $\exp(-(u-\xi)^2/2\sigma^2)/\sqrt{2\pi}\sigma$, $-\infty < u < +\infty$, denoted $N(\xi, \sigma^2)$

Uniform: $1/(2a)$, $-a+\xi < u < +a+\xi$, denoted $U(a, \xi)$

Double exponential: $(\lambda/2)\exp(-\lambda|u-\xi|)$, $-\infty < u < +\infty$, denoted $DE(\lambda, \xi)$

Gamma: $(\lambda^r/\Gamma(r))(u+\gamma)^{r-1}\exp(-\lambda(u+\gamma))$, $u > -\gamma$, denoted $\Gamma(\lambda, \gamma)$ (we consider r fixed)

Exponential: $\lambda\exp(-\lambda(u+\gamma))$, $u > -\gamma$, denoted $E(\lambda, \gamma)$ (gamma distribution with $r = 1$)

To calculate the asymptotic significance levels for the W test, $\varepsilon_W(\theta)$, (see section 3) for the symmetric situations when $p_1(\theta) = 1/2$, we need $p_2(\theta)$ and $p_3(\theta)$. In this case one can show that $p_3(\theta) = p_2(1/\theta)$, so we give only $p_2(\theta)$.

When the two samples are generated from $N(\xi, \sigma^2)$ and $N(\eta, \tau^2)$ respectively, the parameter $\theta = \sigma^2/\tau^2$. Under H: $\xi = \eta$ we obtain

$$p_2(\theta) = \frac{1}{4} + \frac{1}{2\pi}\arcsin\left(\frac{\theta}{1+\theta}\right)$$

When the two samples are generated from $U(a, \xi)$ and $U(b, \eta)$ respectively, the parameter $\theta = (a/b)^2$. Under H: $\xi = \eta$ we obtain

$$p_2(\theta) = \begin{cases} \frac{1}{4} + \frac{1}{12}\theta & \text{if } \theta < 1 \\ \frac{1}{2} - \frac{1}{6}\frac{1}{\sqrt{\theta}} & \text{if } \theta \geq 1 \end{cases}$$

When the two samples are generated from $DE(\lambda, \xi)$ and $DE(\mu, \eta)$ respectively, the parameter $\theta = (\mu/\lambda)^2$. Under H: $\xi = \eta$ we obtain

$$p_2(\theta) = \frac{1}{2}\frac{\sqrt{\theta}}{1+\sqrt{\theta}} + \frac{1}{4}\frac{1}{1+2\sqrt{\theta}}$$

When the two samples are generated from $\Gamma(\lambda, \gamma_1)$ and $\Gamma(\mu, \gamma_2)$ respectively, the parameter $\theta = (\mu/\lambda)^2$. The exponential distribution is a special case of the Γ-distribution with $r = 1$.

ACKNOWLEDGEMENT

We wish to thank Lars Walløe for helpful discussions and valuable comments on the manuscript.

REFERENCES

Best, D.J. & Rayner, J.C.W. (1987) Welch's Approximate Solution for the Behrens-Fisher Problem. *Technometrics* **29**, 205-210.

Birtwistle, G.M., Dahl, O.J., Myhrhaug, B. & Nygaard, K. (1983) *Simula begin.* Chartwell-Bratt Ltd, Bromley.

Bratley, P., Fox, B.L., & Schrage, L.E. (1983) *A guide to simulation.* Springer-Verlag, New York.

Fenstad, G.U. (1983) A comparison between the U and V tests in the Behrens-Fisher problem. *Biometrika* **70**, 300-302.

Lehmann, E.L. (1975) *Nonparametrics: Statistical methods based on ranks.* Holden-Day, Inc., San Francisco.

Scheffé, H. (1943) On solutions of the Behrens-Fisher problem, based on the t-distribution. *Ann. Math. Stat.* **19**, 35-44.

Scheffé, H. (1970) Practical solutions of the Behrens-Fisher problem. *J. Amer. Stat. Assoc.* **65**, 1501-1508.

Skovlund, E. & Walløe, L. (1991) Sequential or fixed sample trial design? A case study by stochastic simulation. *J. Clin. Epidemiol.* **44**, 265-272.

Wang, Y.Y. (1971) Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem. *J. Amer. Stat. Assoc.* **66**, 605-608.

Welch, B. L. (1937) The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

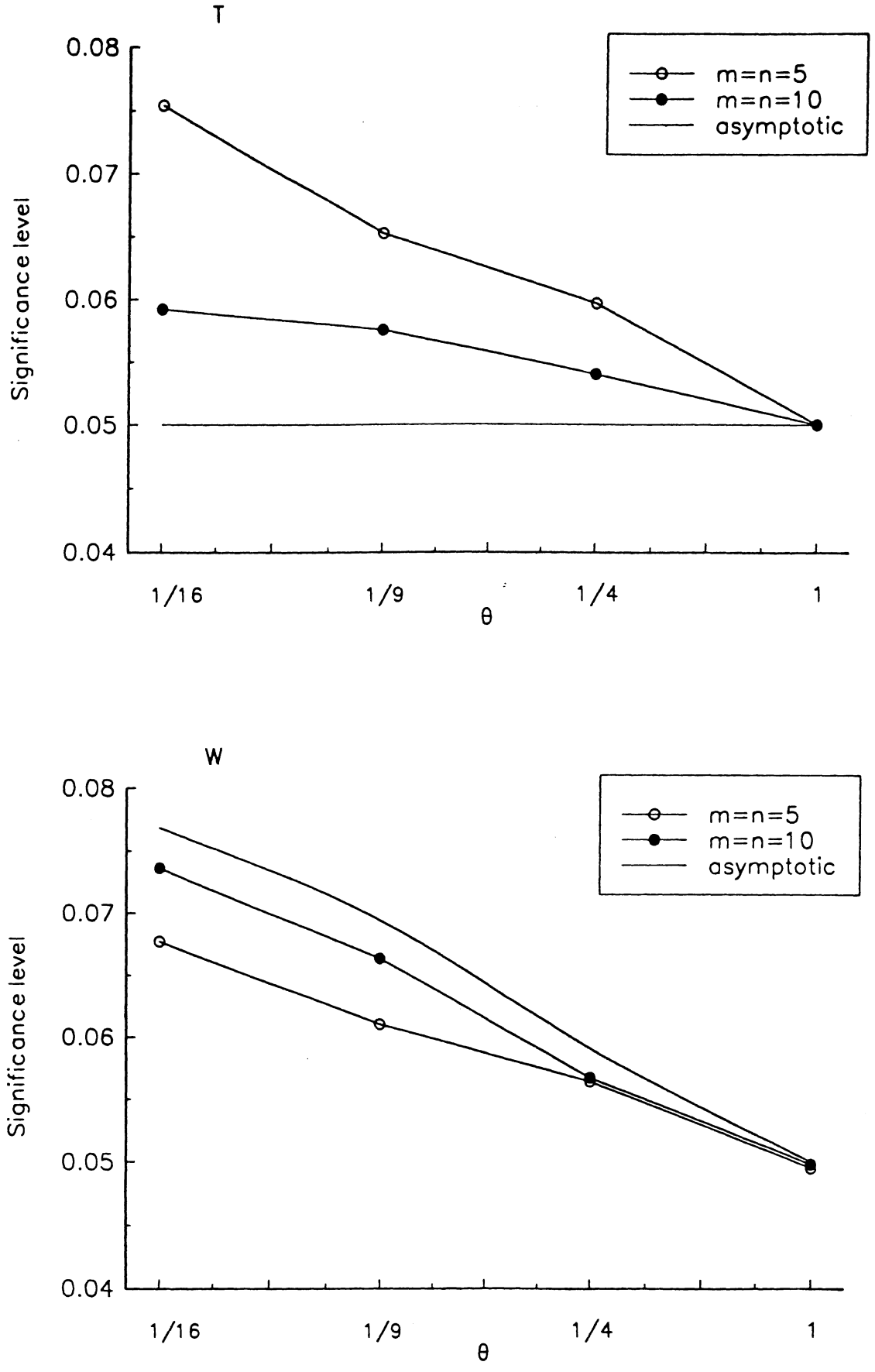

Fig. 1. The upper panel shows simulated significance levels for the T test when $m = n$ and the observations are normal. The asymptotic level is also shown. The lower panel shows asymptotic and simulated significance levels for the W test when $m = n$ and the observations are normal. The nominal level is $\varepsilon = 0.05$.

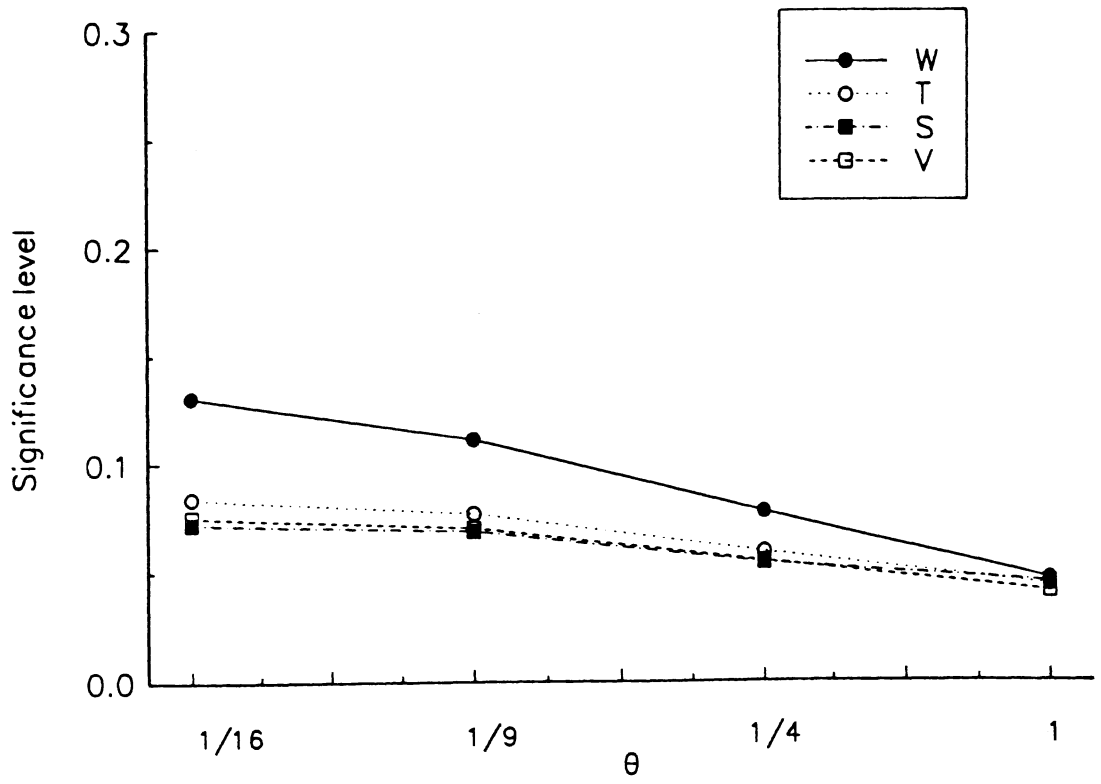

Fig.2. Simulated significance levels of the W, T, S, and V tests when $m = n = 10$ and the observations are generated from a gamma distribution. The nominal level is $\epsilon = 0.05$.

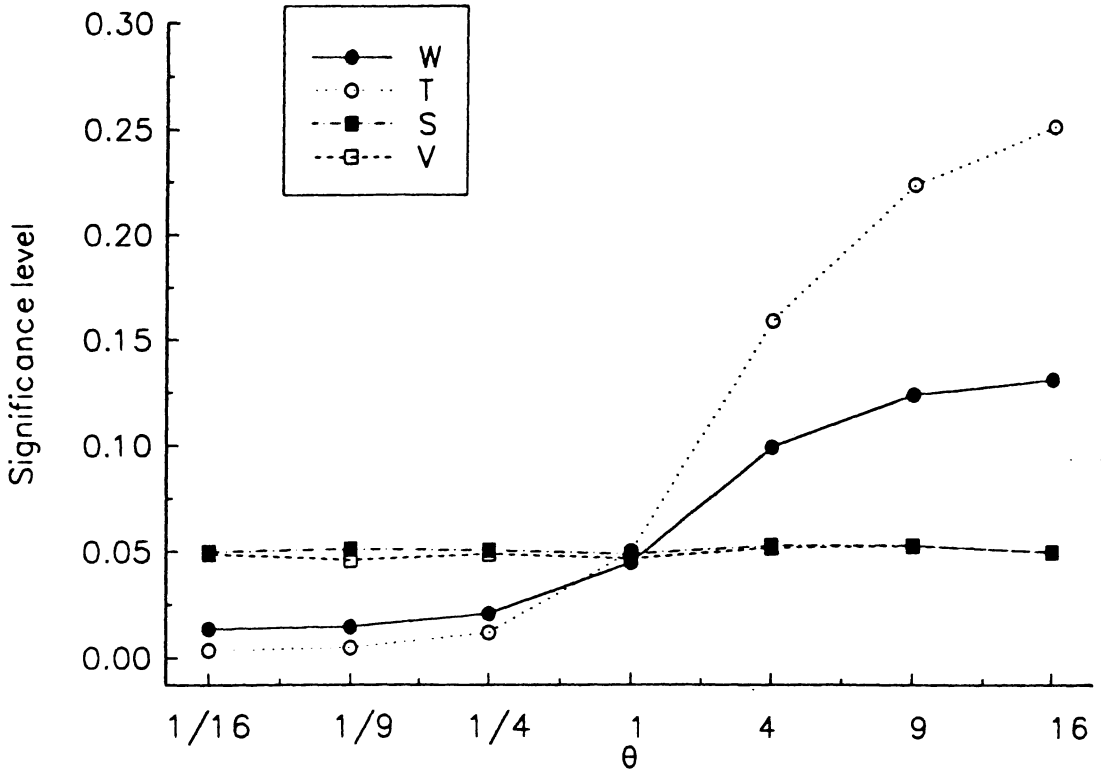

Fig.3. Simulated significance levels of the W, T, S, and V tests when $m = 5$ and $n = 15$ and the observations are normal. The nominal level is $\epsilon = 0.05$.

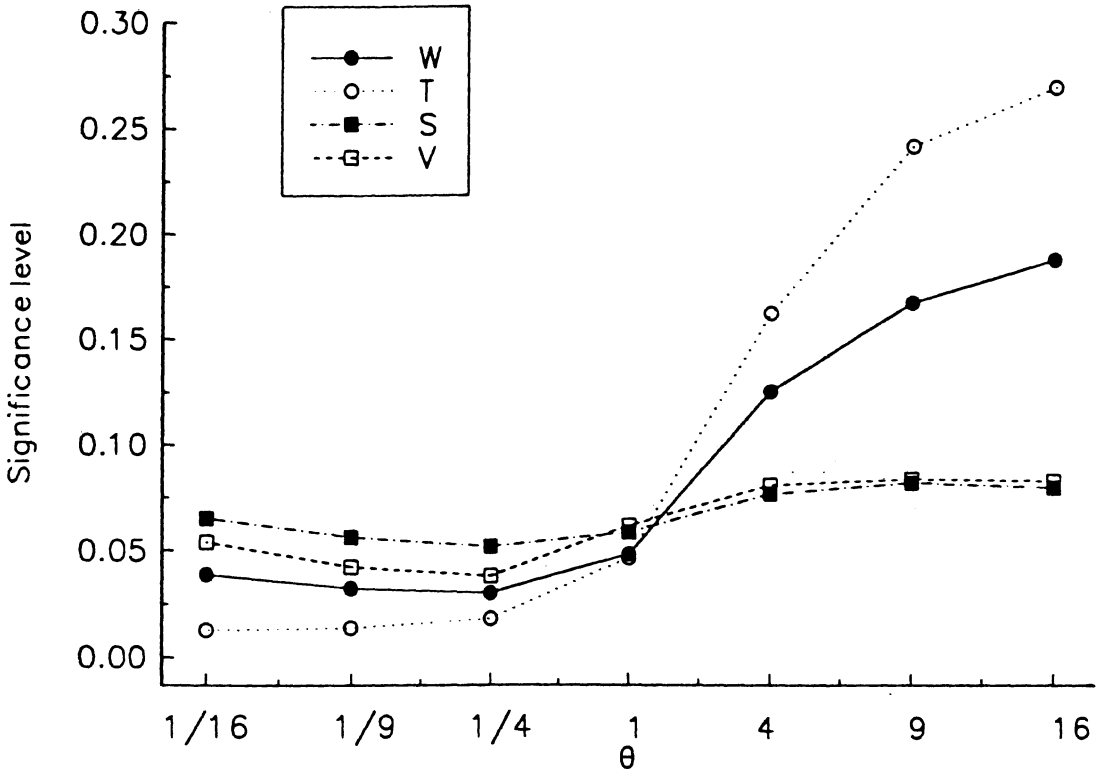

Fig.4. Simulated significance levels of the W, T, S, and V tests when $m = 5$ and $n = 15$ and the observations are gamma distributed. The nominal level is $\varepsilon = 0.05$.

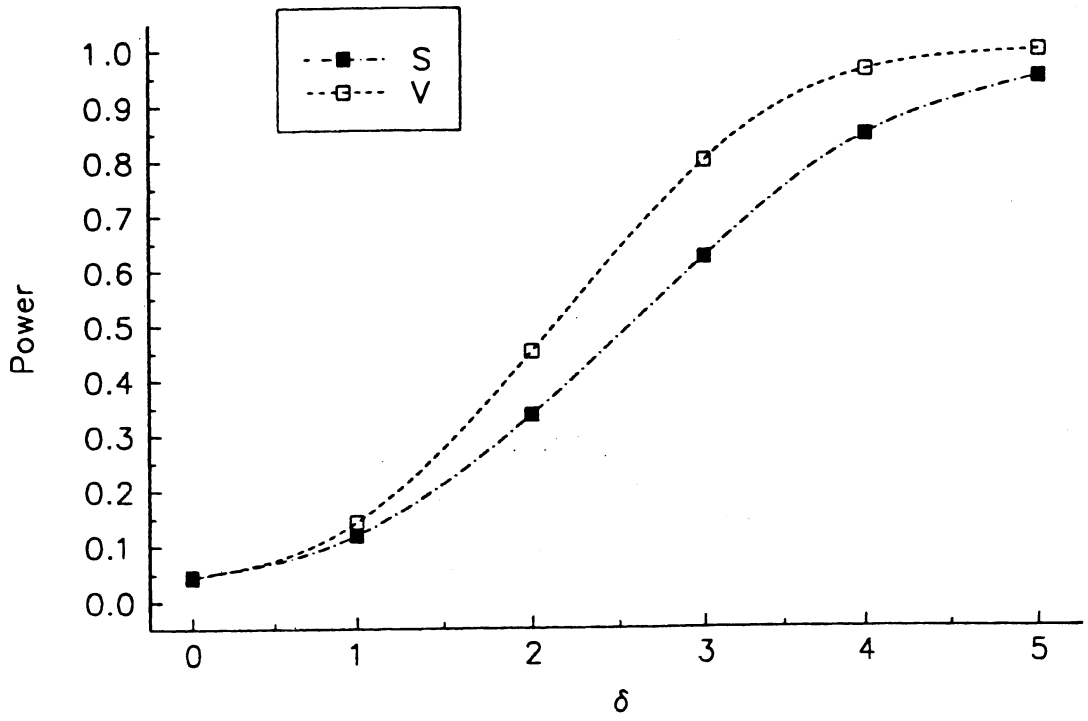

Fig. 5. Simulated power functions of the S and V tests when the observations are normal, $\theta = 1/9$, and $m = 5$ and $n = 15$.